Automatic medical record synthesis using generative AI models: a review*

Carmen Sogan[1,*,†], Ida Tognisse[1,*,†], Pélagie Houngue[1,*,†], Hénoc Soude[1], Abel Yeni M'PO[1] and Jules Degila[1]

[1] *Institute of mathematics and physical sciences (IMSP), Republic of Benin, Dangbo*

Abstract

Electronic Medical Records (EMRs) is a major challenge for modern healthcare systems. Although EMRs centralize information that is essential for patient care, their increasing complexity is facing challenges for healthcare professionals, particularly in terms of time and accuracy of analysis. Generative artificial intelligence (AI) models, such as those based on transformative architectures (e.g. GPT), offer innovative solutions for automating EMR synthesis, reducing clinicians' cognitive load and improving decision-making processes. However, challenges remain, including biases in training data, textual hallucinations and ethical and confidentiality issues. This paper reviews current research on the use of generative AI models for EMR synthesis, assessing their benefits and limitations. It explores solutions to enhance their reliability and acceptability, including standardized methodologies, bias reduction, and better integration of ethical concerns. Finally, it highlights future directions for improving these technologies and their adoption in clinical practice.

Keywords

Electronic Medical Records, generative artificial, synthesis, transformative architectures

1. Introduction

Electronic Medical Records (EMR) centralize essential patient data, including clinical notes, test results, medical history, prescriptions and consultation summaries. Although EMRs have transformed information management in the medical sector, their sheer volume and diversity make them complex to analyze and synthesize. The diversity of formats (free text, medical images, structured data) and the exponential growth of data make their efficient exploitation particularly challenging for healthcare professionals [1, 2]. They spend a significant proportion of their time sorting, reading and summarizing this information to obtain a clear, usable overview. This information overload can lead to human error and delays in decision-making process, and, in some cases, compromise the quality of care. In a context where rapid and accurate decisions need to be made, the lack of effective tools to synthesize EMRs exacerbates these challenges [3]. Artificial Intelligence (AI) models, particularly generative models such as those based on transform architectures (e.g., GPT), offer innovative perspectives to address these issues. These models can generate synthetic text from unstructured data, making it possible to automatically summarize EMRs, extract key information, and produce summaries tailored to clinicians' needs. Studies have shown that these technologies can lighten the cognitive load of caregivers, reduce the risk of human error and improve the efficiency of the decision-making process [4, 5]. For example, Myers et al [2] have shown that generative AI models sometimes outperform humans in terms of speed and consistency when summarizing hospital reports. Moreover, Shing et al [3], have highlighted the ability of these models to improve access to essential information while reducing the

*CITA2024 – Emerging Technologies and Sustainable Agriculture, 26-28 June 2025, Cotonou, Benin,

* Corresponding author.

† These authors contributed equally.

sogancarmen1@gmail.com (C. Sogan); ida.tognisse@imsp-uac.org (I. Tognisse); pelagie.houngue@imsp-uac.org (P. Houngue); abel.yenimpo@imsp-uac.org (A. Y. M'PO)

0000-0003-4688-9178 (J. Degila)

administrative burden on clinicians. However, despite these advances, major challenges remain. Models can be sensitive to biases inherent in training data, leading to inconsistent or erroneous results in critical clinical settings [6]. Nguyen et al [7] have warned of the potential impact of these biases on automated medical decisions, increasing the risks to patients.

In addition, textual hallucinations where the model generates incorrect or unsubstantiated information remain a major concern [10]. Simmons et al [8] have proposed mechanisms such as validation filters to alleviate these problems. At the same time, ethical issues, data confidentiality and acceptability to healthcare professionals are holding back the adoption of these technologies in clinical practice. The work of Goodman et al [1] has highlighted the importance of robust regulation and ethical standards to ensure the safe and fair use of AI models in the medical field. Faced with these challenges, a thorough assessment of existing research is crucial to better understand the potential and limitations of AI models applied to EMR synthesis.

This review aims to answer the following questions:

- What are the main benefits and challenges associated with their use?
- How do these models influence clinical practice and decision-making in healthcare?
- What are the future directions for improving their reliability, safety and acceptability?
- What approaches based on AI models have been explored for DME (Electronic Medical Record) synthesis?

By exploring these dimensions, this analysis intends to provide enlightening perspectives on the future of AI in healthcare systems while highlighting the steps needed to overcome current obstacles. In this article, we begin with a literature review, before presenting the methodology used and the results of the bibliometric study. We then provide a summary table of the ten best articles identified, before concluding.

2. Literature review

The use of artificial intelligence (AI) models in medical record synthesis represents a significant advance in the field of digital health. Several researchers have conducted surveys to gain a deeper understanding of AI's contribution to medical record synthesis. Xie H and al [9] have developed a bibliometric analysis of Australian literature (1991-2022) on electronic medical records research trends and patterns. Overall, they reveal the impact of EMRs on the digital transformation of the healthcare system in the face of demographic and pandemic challenges. The study serves both as academic mapping and as a decision- support tool for public policy. Similarly, Ananda Haris and al [10] conducted a bibliometric analysis of the acceptance and adoption of electronic medical records (EMRs). This study provides a bibliometric analysis of research on the adoption of electronic medical and health records (EMR/EHR). The study also reveals gaps in research, notably on cybersecurity and user satisfaction, while highlighting the potential of these systems to optimize care on a global scale. Jeena Joseph Jr and al [11] set out to map the landscape of electronic medical records and health information exchange using bibliometric analysis and visualization. Their findings reveal major challenges such as interoperability, data privacy and the integration of emerging technologies like AI and blockchain, while also highlighting research gaps and innovation opportunities to improve the efficiency and quality of care. Elsewhere Yaojue Xie et al [12] have explored the evolution of artificial intelligence in healthcare: a 30-year bibliometric study. Their analysis reveals a gradual increase in publications since the 1990s, with a marked

acceleration after 2015, linked to the emergence of deep learning techniques. The study also identifies the clinical areas most influenced by AI, such as assisted diagnosis and medical imaging analysis. Saadat M. Alhashmi et al [13] carried out a bibliometric analysis of the application of artificial intelligence in healthcare. They highlight the most prolific countries, institutions, and authors in this field, while highlighting the evolution of keywords and themes over time. Their study also maps international collaborations and suggests a growing need for standardization and ethics in medical AI research. Feng chen et al [14] carry out a systematic review of biases in artificial intelligence (AI) models based on electronic medical records (EHR). The authors highlight the need for standardized norms and real-life testing to ensure fair applications of AI in the medical field. Abdul Khalique Shaikh et al [15] carry out a bibliometric analysis to study the adoption of artificial intelligence (AI) applications in the e-health sector. They analyze research trends over a 25-year period (1996-2021) using the Scopus database, highlighting the most influential authors, institutions, countries, journals and keywords. Evrim Özmen et al [16] carry out a retrospective bibliometric analysis to assess the role of machine learning algorithms in the diagnosis of sepsis. Silvana Secinaro et al [17] carry out a structured literature review on the role of artificial intelligence (AI) in the healthcare sector. Their findings highlight the potential of AI to improve diagnosis, personalized treatment and patient data management. Elham Asgari et al [18] examine the impact of electronic medical records (EHR) on clinicians' cognitive load and burnout.

While much of the work focuses on the evolution, adoption, biases, cybersecurity and overall impact of AI and EMRs in healthcare, unlike, this systematic review paper would focus on the ability of AI models to synthesize and generate medical information. It would provide a targeted overview of algorithmic approaches, corpora used, evaluation metrics, and concrete applications in clinical settings, while highlighting methodological limitations, clinical validation needs, and avenues for improvement, thus contributing to an original and operationalization-oriented way to the existing literature.

3. Methodology

This applied research also included a bibliometric study. The data for this research was collected from the Scopus database, comprising 691 documents published between 2020 and 2025. Among the different types of documents available in Scopus, several relevant categories were considered for this study, including Articles, Journals, Proceeding Papers, Review Articles, as well as other types specific to this database. These documents were selected based on their relevance to the research topic. We have decided to consider only those studies carried out and published between 2020 and 2025, as it is during this period that transformative and generative models in the medical field appear and gain in importance. We also note an explosion in publications following the arrival of models such as GPT-3 (2020) Tom Brown et al [19] and especially ChatGPT/GPT-4 (2022-2023) Open AI GPT-4 Technical Report (2023) [20], Kung et al. (2023) [21]. The choice of a bibliometric approach is explained by the desire to obtain a quantitative overview of scientific production on the use of generative AI models for the synthesis of electronic medical records (EMRs), between 2020 and 2025. Unlike a systematic review, which would have enabled an in-depth analysis of the content of the selected studies. The specifics of the data collection phase are summarized in Table 1.

Attribut	Value
Search chain	(Synthesis OR generation) AND ("Medical Record" OR "medical file" OR "electronic health records" OR "health records") AND ("artificial intelligence" OR "deep learning" OR "machine Learning" OR "automatique Learning")
Database	Scopus
Year	2020-2025

Search date	20 April 2025
Document N°	691

Two main tools were used to analyze the data collected: VOSviewer (version 1.6.18), for visualizing co- occurrence and collaboration networks, and RStudio's Bibliometric library, enabling detailed bibliometric analyses. A comparative systematic review was carried out on the 10 best articles that we had selected using the bibliometric study based on the highest to lowest number of citations.

4. Bibliometric study result

4.1. Annual scientific production

Figure 1 illustrates annual scientific output. The graph was generated using the Bibliometric package under R. The graph shows exponential growth from 2023 onwards, with a peak in 2024. Prior to 2020, the field was little explored, but the rise of AI technologies and their application to medical records has probably stimulated researchers' interest. This recent increase confirms that the subject is booming and underlines its relevance to current research.

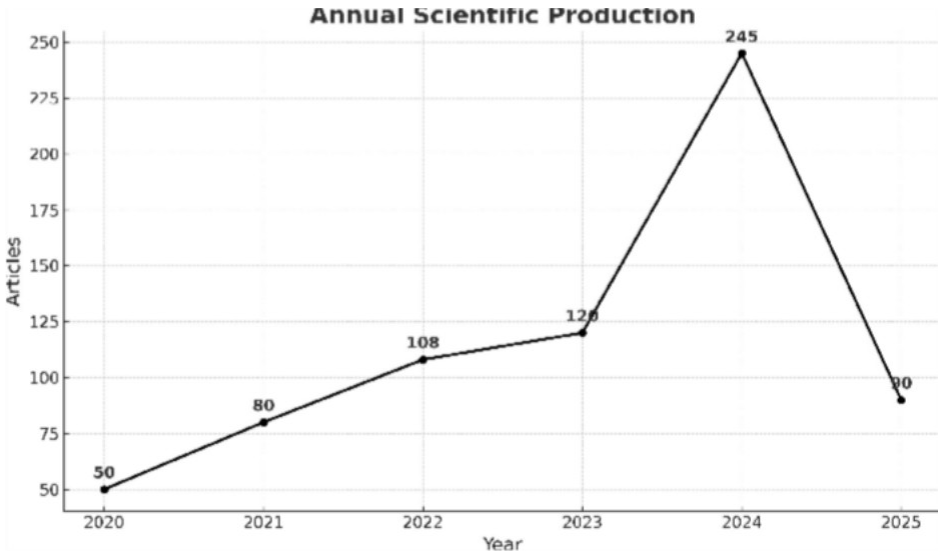

Figure 1: Annual scientific production

4.2. Most relevant sources

Table 1 shows that "JMIR MEDICAL INFORMATICS" is the dominant source in this corpus, followed by journals and conferences focusing on biomedical informatics and artificial intelligence in medicine.

Table 1: The most frequently cited sources

Sources	Articles
JMIR MEDICAL INFORMATICS	16
JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	15
JOURNAL OF BIOMEDICAL INFORMATICS	14
LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS)	14
LECTURE NOTES IN NETWORKS AND SYSTEMS	14
JOURNAL OF MEDICAL INTERNET RESEARCH	13
IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS	11
BMC MEDICAL INFORMATICS AND DECISION MAKING	10
STUDIES IN HEALTH TECHNOLOGY AND INFORMATICS	10
COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE	9

Although some sources have a limited number of documents, their specialization can make them crucial for niche topics. Overall, this distribution highlights the most influential publications for research in fields related to informatics and health.

4.3. Scientific output by country

Figure 2 illustrates the scientific output by country. The graph was generated using the Bibliometric package under R. The USA dominates scientific output in AI applied to medical records, with a strong acceleration after 2022. China shows rapid growth since 2021, indicating increased investment. India, the UK and Australia show more modest but growing contributions, especially after 2020. These dynamics reflect a global rise in interest in healthcare AI, with development concentrated in certain leading countries.

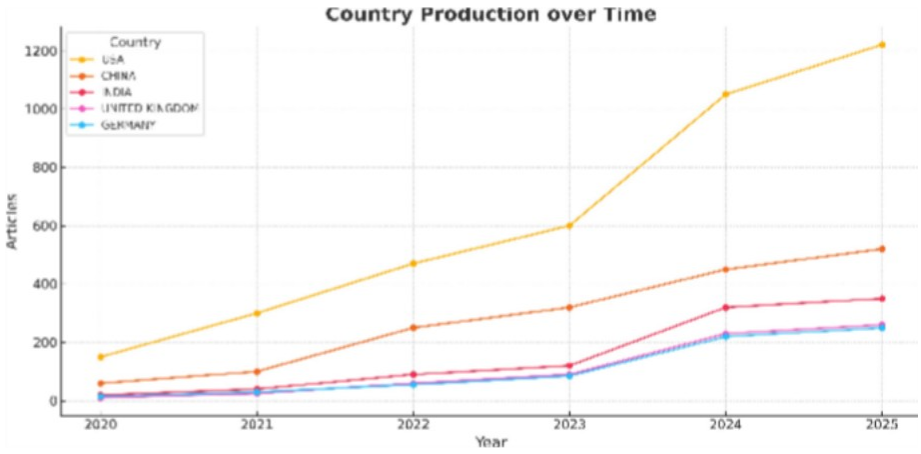

Figure 2: Scientific output by country

4.4. Most popular countries

Table 2 illustrates the most popular countries. The United States comes out on top with 2214 citations, underlining its strong influence in this field. Germany, China, the UK, Spain, India and Korea follow at a distance with notable contributions. Canada, Italy and Pakistan show a more modest impact. This reveals that scientific recognition is mainly concentrated in countries with advanced research resources.

Table 2: Most frequently cited countries

Country	TC	Average Article Citations
USA	2214	15.00
GERMANY	944	41.00
CHINA	451	5.50
UNITED KINGDOM	335	12.00
SPAIN	294	16.30
INDIA	266	4.40
KOREA	220	15.70
CANADA	179	12.80
ITALY	176	13.50
PAKISTAN	166	23.60

4.5. Co-citation network

This co-citation map illustrates the publications most frequently referenced together within the analyzed corpus. It reveals several thematic clusters: a central group focused on the application of Artificial Intelligence in healthcare (notably around Esteban C. and Choi E.), a green cluster centered on foundational work related to generative adversarial networks (GANs), and another cluster linked to large language models, including the GPT-4 technical report. Additionally, the map highlights publications addressing ethical considerations and fairness in AI. The size of each name indicates its influence in the field, while the proximity between them reflects strong bibliographic connections, helping to uncover the main theoretical underpinnings of the research landscape.

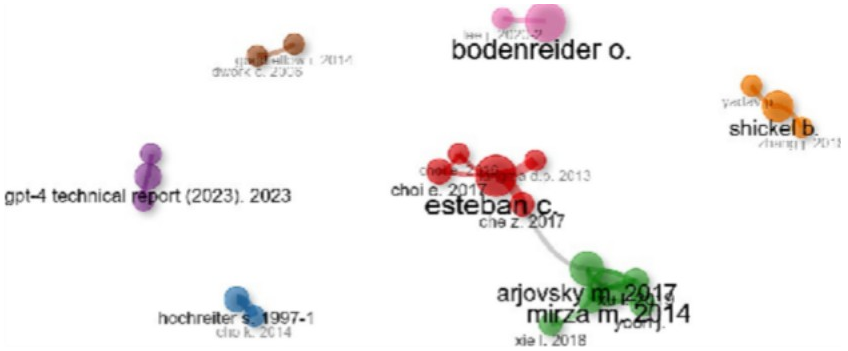

Figure 3: Co-citation network

4.6. Keyword co-occurrence network

The keyword co-occurrence network reveals a threefold thematic structure within AI-driven health research. This structure encompasses a technological dimension (e.g., electronic health records, machine learning), a clinical dimension (e.g., diagnosis, medical data), and a humanistic-conceptual dimension (e.g., terms like “human” and “article”), all intricately connected. Artificial intelligence and the human factor emerge as central elements, symbolizing the convergence of algorithmic innovation, real-world clinical implementation, and ethical considerations. Notably, the frequent appearance of the term “article” underscores the enduring role of scholarly output in shaping this interdisciplinary dialogue, where technical, clinical, and ethical domains intersect.

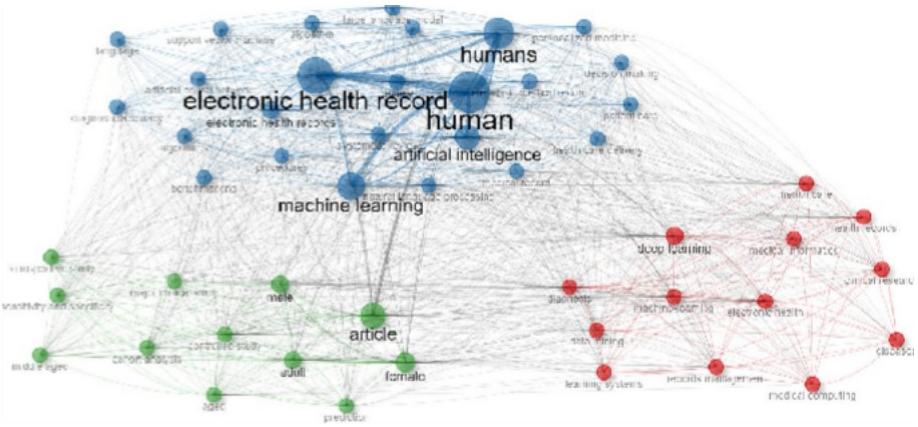

Figure 4: Keyword co-occurrence network

4.7. Trend analysis results

The analysis of thematic trends illustrates the evolution of major scientific terms from the 1990s to the present day. The analysis reveals an evolution of research between 2020 and 2024 centered on three major axes: advanced AI technologies (adversarial network, transfer learning, natural language processing, data mining), medical applications (electronic health records, standardized medical nomenclature, genotyping), and technical infrastructures (IT security, analysis software, learning systems). Recurring terms such as “artificial intelligence”, “human” and “article” underline the interdependence between algorithmic innovation, clinical needs and scientific production, reflecting an increasingly integrated research environment where technology serves as a bridge between medical theory and practice.

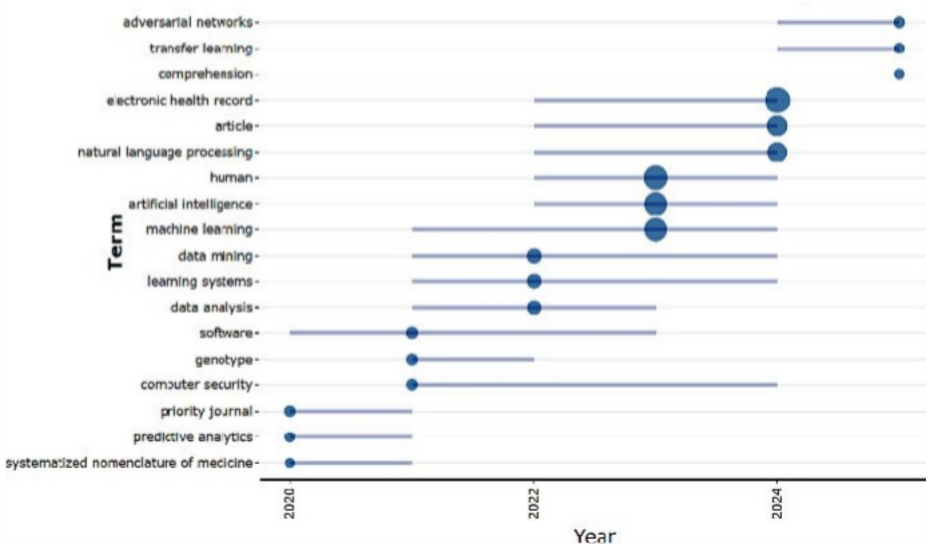

Figure 5: Trend analysis results

4.8. Analysis of word frequency over time

Analysis of the cumulative frequency of words over time highlights the evolution of dominant themes in scientific publications. Analysis of cumulative occurrences highlights the following key concepts: articles, artificial intelligence, deep learning, electronic medical record, woman, human, human being, machine learning, human and automatic natural language processing. These terms reflect a strong focus on AI technologies applied to healthcare, with particular attention paid to electronic medical data, gender differences (female/male), and human aspects, while integrating

advanced techniques such as deep learning and NLP, illustrating the interdisciplinarity of contemporary digital health research.

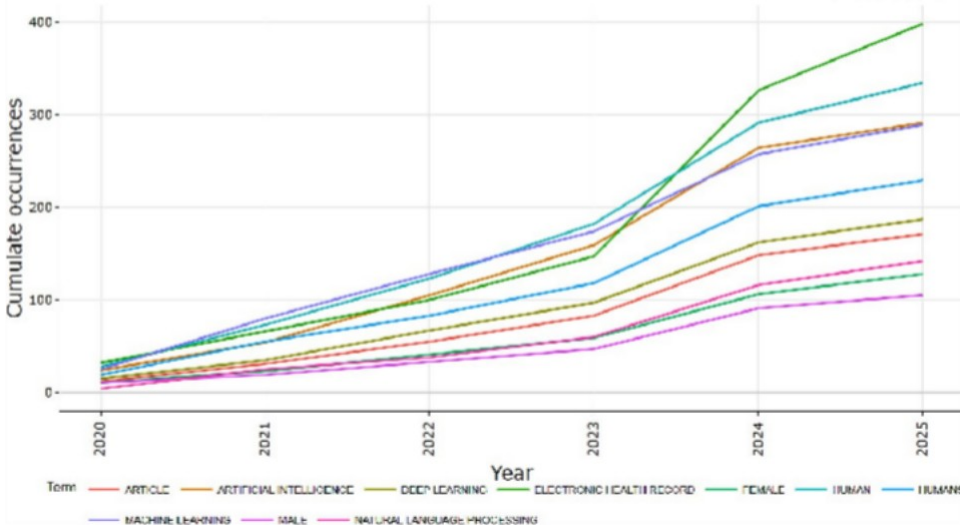

Figure 6: Analysis of word frequency over times

4.9. Local impact of authors using the H index

The table on authors' local impact, measured by the H-index, highlights a hierarchy among scientific contributors in the field of automatic medical record synthesis with generative AI. We note that all three authors have an H-index of 5, five authors have 4 and two authors have 3. These results show that a handful of authors dominate this emerging field, reflecting a concentration of research efforts.

Table 3: Most frequently cited documents

Author	h_index	g_index	m_index	TC	NP	PY_start
LI Y	5	6	1.000	149	6	2021
WANG F	5	6	0.833	276	6	2020
WANG Y	5	9	0.833	96	10	2020
KIM J	4	5	0.800	30	5	2021
LIU H	4	5	0.800	48	5	2021
LIU Y	4	6	0.800	43	8	2021
XIE F	4	4	0.667	115	4	2020
ZHAO J	4	5	0.800	42	5	2021
ANGULO C	3	3	0.600	42	3	2021
BENEDUM CM	3	3	0.750	50	3	2022

4.10. The best papers

The ten selected articles collectively explore the diverse and evolving applications of artificial intelligence, particularly deep learning models—in the healthcare sector, with a focus on electronic medical records (EMRs), real-world data, mental health, and ethical considerations. Chang Su et al. review the use of deep learning for mental health outcome prediction, highlighting its superior performance compared to traditional techniques, though challenges persist in validating results due to anonymized social network data. Szu-Wei Cheng et al. examine the current and potential uses of ChatGPT in psychiatry, recognizing its value in administrative and communicative tasks but noting its limited clinical reliability and ethical concerns. Fang Liu et al. offer a comprehensive overview of real-world data (RWD) sources and their analytical frameworks, underlining the need for rigorous preprocessing to ensure reliability in clinical trials. Yinan Huang et al. focus on machine learning algorithms predicting hospital readmissions, identifying neural networks and decision trees as the most effective, albeit with issues related to validation consistency and data standardization. Junyu Luo et al. propose HiTANet, a temporal attention network that captures

non-stationary disease progression, outperforming traditional static models, although its sensitivity to EHR data quality remains a limitation. Feng Xie et al. intro- duce AutoScore, a machine-learning- based tool that generates interpretable clinical scoring systems, achieving strong predictive performance with fewer variables and improved usability. Isotta Landi et al. develop a deep unsupervised learning framework (ConvAE) for large-scale EHR-based patient stratification, effectively identifying meaningful subgroups in diseases like Parkinson's and diabetes, though semantic depth and generalizability require further work. Gaurav Dhiman et al. present a hybrid CNN model for tumor identification in medical imaging, significantly improving information extraction, but with concerns over the quality and semantic coherence of pseudo-data. Ping Wang et al. build TREQS, a model translating natural language questions into SQL queries to simplify EMR access for healthcare professionals, demonstrating high accuracy and robustness to typos and abbreviations, yet struggling with complex queries. Finally, Karan Bhanot et al. address fairness in synthetic health data by proposing equity metrics and applying them across datasets like MIMIC-III, revealing systematic under representation of subgroups such as elderly or minority populations. Together, these studies emphasize the transformative potential of generative and predictive AI models in digital health, while underlining the importance of model interpretability, data diversity, ethical safeguards, and validation in real-world clinical environments.

Table 4: Best Papers

Author(s)	Objectives	Research Questions	Theory/Context	Methodology	Results	Limits	Future Work	Research Direction
Chang Su and al.[22]	review existing research on applications of deep learning algorithms in mental health outcomes research	How are deep learning algorithms being applied to improve the diagnosis and treatment of mental health conditions? What is the different categories of data used in these studies, and how do these data influence the results?	The background to this research is based on the growing recognition that mental illnesses such as depression are common and affect the physical health of individuals. The study explores how artificial intelligence, particularly deep learning, can assist mental health professionals in their clinical decisions.	systematic literature review, guided by PRISMA guidelines.	Deep learning models often achieve higher predictive accuracy than traditional machine learning techniques.	Difficulty in validating results due to the anonymization of social network data. Lack of in-depth analysis of users' social network struc- tures. The need for greater integration of domain knowledge to improve the validity of results.	Integrating deep learning models with therapeutic interventions	Exploring social networks as a potential tool for studying mental health problems. Improving the translatability of results into practical innovations for real time interventions. Incorporating health-specific knowledge to overcome data limitations.

Fang Liu and al.[23]	Provide an overview of Real World Data (RWD) types and sources, as well as models and approaches for their use and analysis.	How can RWD be used to generate real evidence and answer questions that cannot be addressed by traditional clinical trials?	The U . S . FDA's definition of RWD, which includes patient health and healthcare delivery data collected from a variety of sources.	Examination of different sources of RWD (such as electronic health records and data from wearable devices)	Real-world data offer great potential for designing and conducting validating clinical trials and can help answer important questions. However, the complexity of the data requires the development of appropriate and rigorous analysis techniques.	Challenges include the variability and complexity of the data, requiring significant effort for preprocessing and analysis.	The need for further research into the explicability and interpretability of machine learning models, as well as methods to guarantee the reproducibility and replicability of results.	Exploring new techniques for the treatment of RWD while considering ethical issues and the need to maintain scientific rigor in the use of data.
Junyu Luo and al.[24]	Proposed a risk prediction model capable of better capturing temporal information in electronic health records (EHR) through a hierarchical, time-aware attention net-	How can we model disease progress in anon-stationary, dynamic way? What are the key time steps for patient-specific disease prediction?	Current approaches assume stationary disease progression, which is not the case in reality. This work criticizes this assumption and proposes a model that incorporates mechanisms to reflect the	Implementation, simulation, case study	proposal of an innovative approach to deal with the limitations of previous models in taking account of temporal information	The method may be sensitive to the quality of the EHR data, and results may vary depending on the characteristics of the datasets.	Explore other attention mechanisms, integrate additional data from different sources, and extend the model to other pathologies.	Future research could focus on improving the model's robustness in the face of incomplete or unbalanced data, as well as on methods for integrating external medical knowledge.

	work, called HiTANet.		decision making process.					
Isotta Landi and al.[25]	Propose a general frame- work based on unsupervised deep learning to exploit elec- tronic health records (EHRs) on a large scale and scalable, without the need for manual feature engineering or supervision.	How can dis- ease subtypes be extracted automati- cally and efficiently from large, het- erogeneous EHR databases? Can we design an unsu- pervised method capable of pro- ducing clinically meaningful group- ings? Finally, do the representations generated by Con- vAE enable better performance in terms of patient clustering, com- pared with existing methods?	EHRs, although heterogeneous, contain valuable information on patients' health tra- jectories. However, their use for patient stratification remains limited by their complexity, volume and vari- able quality. Many diseases, such as type 2 diabetes, Parkinson's disease or Alzhe imer's, present a high degree of clinical heterogeneity. This complexity makes them dif- ficult to model using conventional approaches.	Implementa- tion, simulation	ConvAE sig- nificantly outperforms existing methods such as Deep Patient and other ap- proaches based on linear repre- sentations. The model identified several clinically rele- vant subgroups for complex diseases such as type 2 dia- betes, Parkin- son's and Alzheimer's.	the quality of EHR data may introduce noise, bias or uninforma- tive concepts; the model has only been tested on certain complex diseases and does not cover all possible conditions; subtype analysis is based solely on concept frequency, with no in- depth semantic consideration	introduce multi-level clustering for finer stratifica- tion, test the model on EHRs from other institutions to validate its generalizabil- ity, and exploit the subtypes discovered as labels for supervised prediction models. integrate other data sources, such asgen omics, to enhance	This work paves the way for a new generation of integrated clinical systems, capable of rep- resenting each patient with a single, robust vector that can be used for a variety of medical tasks.

							disease characterizatio n and improve the predictive capabilities of their system.	
Feng Xie and al.[26]	propose and demonstrate the effectiveness of AutoScore, an automatic clinical score generator based on ma- chine learning, designed to produce inter- pretable score models that can be easily used in a variety of medical contexts	how to automate the process of generating inter- pretable clinical scores from elec- tronic health record data? how does this system compare with con- ventional statistical approaches in terms of predictive performance, sim- plicity and clinical applicability?	The work is based on the growing need for transpar- ent, robust and easily applicable clinical prediction tools. Traditional scoring models are often built manually, which limits their adaptability. AutoScore fits in with the growing use of electronic medical records and the rise of artificial intelligence in healthcare.	Implementa- tion, simulation, case study	The nine- variable AutoScore model achieved an area under the curve (AUC) of 0.780, comparable to more complex regression models. It demonstrated good calibration and used fewer variables, while remaining highly interpretable. The system offers an optimal compromise between predictive performance and simplicity. The authors draw on	Limitations include the use of a single dataset based on routine variables, the absence of some relevant variable s (such as specific intensive care), the retrospec- tive nature of the study, and the need for prospective validation.	The authors suggest the integration of more advanced techniques into AutoScore modules, adaptation to other contexts (smaller co- horts, other fields such as finance or justice), and prospective validation of its actual clinical efficacy.	Future research should focus on modular enhancement of AutoScore with state-of- the-art algorithms, adaptation to diverse data types, and setting standards for interpretable, scalable and automatically generated clinical scores to support medical decisions.

					previous studies of clinical score models such as SOFA, NEWS, or HEART score and compare their approach with well- established techniques such as step- wise regression, LASSO, and random forests.			
Yinan Huang and al.[27]	synthesize the literature on machine learning (ML) methods used to predict hospital read- missions in the USA, focusing on model performance and the types of algorithms employed.	Which machine learning (ML) algorithms are most commonly used to predict hospital readmissions in the USA? How does the performance of ML models, in particu- lar the AUC (area under the curve), vary according to the different meth- ods? What are	The context is based on the importance of reducing hospi- tal readmissions to improve quality of care and cut costs, notably as part of the Hospital Read- mission Reduction Program (HRRP) in the United States.	The methodology followed the PRISMA-ScR and CHARMS guidelines, including a systematic search of the PUBMED, MEDLINE and EMBASE databases from 2015 to 2019, with qualitative and	The results showed that methods based on decision trees, neural networks and regularized logistic regres- sion were the most widely used, with variable performance (median AUC of 0.68).	Limitations include the lack of stan- dardization in validation methods, the absence of comparison with traditional statistical models, and the paucity of studies	Future work should com- pare the performance of ML algorithms, incorporate more clinical variables and improve external validation methods.	Future research could explore the application of ML to specific populations and the use of advanced techniques such as natural language processing to extract unstructured data

		the main data sources (electronic medical records, adminis- trative databases, etc.) used in these studies? What are the challenges and limitations associated with applying ML to predict hospital readmissions? How are clinical and socio-demographic variables integrated into these predictive models?		quantitative assessment of the selected studies.		reporting metrics other than AUC.		
Gaurav Dhiman and al.[28]	propose a hybrid model based on machine learning for tumor identification in medical image processing.	How can we improve the joint extraction of discrete attributes of tumor-related medi- cal events, such as primary site and	The context of this research is based on natural language processing (NLP) and the extraction of medical informa-	Implementation, simulation, case study	The results show that the proposed model achieves an F1 score of 73.52 on the CCKS2020 dataset, ranking third in	Limitations include the strong randomization of the pseudo-data generation algorithm,	The authors plan to improve the pseudo-data generation algorithm based on semantic	Future research could explore the integration of pre-trained language models with less dependence on external resources, as well as the

		tumor size? How can pseudo-data be generated to overcome the lack of annotated data and improve the model's transfer learning capability? How does the proposed model compare with existing methods on the CCKS2019 and CCKS2020 datasets?	tion from electronic medical records. The authors empha- size the importance of t u m o re r e l a t e d events, such as primary site, size and metastasis, and highlight the limitations of existing methods, such as lack of generalizabiliy and reliance on pre-trained models.		the evaluation task. It also out performs the CCMNN method with a significant improvement, particularly for primary tumor size ex- traction (+8.93 on CCKS2019 and +7.51 on CCKS2020).	which can produce data that does not conform to natural semantics, thus affecting model performance. In addition, the model is highly dependent on the quality of the av ailable annotations.	similarity to produce more consistent data. They would also like to extend the application of the model to other types of medical events	optimization of data augmenta- tion techniques for specific medical fields.
Ping Wang and al.[29]	develop a model capable of translat- ing natural language questions on electronic medical records	How to automat- ically generate SQL queries from natural language questions in the medical field? How to handle abbreviations and common typos in medical questions?	The context is based on the in- creasing use of electronic medical records and the need for efficient tools to interrogate this data. Existing approaches, such	Implementa- tion, simulation, case study	The results show that the TREQS model outperforms existing meth- ods in terms of accuracy, particularly for the generation of condition values. It is also robust to	Limitations include depen- dence on the quality of the training data and the difficulty of generalizing the model to	Future work could include extending the model to other medical databases, improving the handling of complex questions and	The research di- rection aims to improve inter- action between healthcare professionals and medical information systems, by developing more

	(EMRs) into SQL queries, to facilitate access to medical information for healthcare professionals without the need for database skills.	How can I improve the accuracy of condition values in generated SQL queries?	as sequence-to-sequence models (Seq2Seq), are suitable but need to be improved to meet the specificities of the medical field, such as abbreviations and technical terms.		questions containing abbreviations or errors, thanks to its value recovery mechanism.	other medical fields without specific adaptation. In addition, the model may encounter difficulties with highly complex or ambiguous questions.	integrating active learning techniques to reduce the need for annotated data.	robust and adaptive models for generating SQL queries from natural language questions.
Szu-Wei Cheng and al.[30]	Explore current and future applications of ChatGPT and GPT technology in psychiatry.	what are the current uses of Chat- GPT in psychiatry? What are its limitations? how can it be ethically and effectively integrated into mental health care in the future?	The theoretical framework is based on the evolution of NLP models towards more contextual and generative approaches with GPT, which offers unique potential in natural language interpretation.	This is a critical and forward-looking review of current GPT usages, enriched with concrete examples, use cases and comparisons with previous approaches.	Results show that ChatGPT is currently useful for administrative tasks, communication between professionals and with patients, and as a writing and research support tool	The main limitations concern ChatGPT's ten- dency to hallu- cinate facts, its lack of clinical reasoning, and its average performance in assessing personality	Future work suggested by the authors includes the integration of empathy capabilities, emotional recognition, detection of signs of mental suffering, and the	In terms of research direction, they call for solid ethical standards and the development of humane, user-friendly AI interfaces capable of co-operating with mental health professionals.

						and suicidal ideation	eventual creation of an automated psychotherapy system.	
Karan Bhanot and al.[31]	study the problem of equity in synthetic health data, proposing metrics to assess the representativeness of subgroups defined by protected attributes (such as age, gender or ethnic origin).	How can we measure equity in synthetic health data to ensure fair representa- tion of protected subgroups ? What inequalities exist in published synthetic datasets, and how do they manifest themselves? How can equity metrics be adapted for temporal data, such as health time series?	The background to the article is based on the challenges of accessing real health data due to privacy laws, which has led to the increasing use of synthetic data. However, such data can reproduce or amplify existing biases, particularly towards minorities. The authors draw on concepts from machine learning equity, such as disparate impact, and adapt them to evaluate	The authors propose two main metrics: "log disparity" to assess the representativeness of subgroups and a metric based on the number of subgroups. Time series to analyze trends. These metrics are applied to three case studies	The results reveal significant biases in thesynthetic data, such as the under representation of older people or ethnic mi- norities. For example, in the MIMIC- III dataset, black people are under represented, while white people are over represented. Temporal metrics also show discrepancies in capturing trends for certain sub- groups.	The authors point out that existing synthetic datasets can introduce significant biases. Certain subgroups defined by protected at- tributes such as age, gender or ethnicity are often underrepresented or misrepresented, which can compromise the fairness of analyses and models	The authors suggest exploring other equity metrics, incorporating equity constraints into synthetic data generation methods (such as conditional GANs), and extending the analysis to other data types and contexts.	Future research should focus on developing methods for generating synthetic data that explicitly incorporate equity criteria, as well as assessing the impact of bias in real- world applications, such as pre- dictive health models.

			synthetic data.	using different datasets (MIMIC-III, ATUS and ASD). Statistical tests and visualizations are used to identify biases.		developed from these data. The disparity measures used reveal that these biases can affect both univariate and multivariate analyses, making conclusions potentially non-generalizable to real data.		

5. Approach: Fair NLP-based synthetic EMR generation pipeline in Benin

In this article, we propose a comprehensive methodological architecture aimed at generating, testing, and validating synthetic Electronic Medical Records (EMRs) from real data from medical platforms. The objective is to:

- preserve patient confidentiality,
- ensure the fairness of generative models,
- and facilitate rigorous clinical validation.

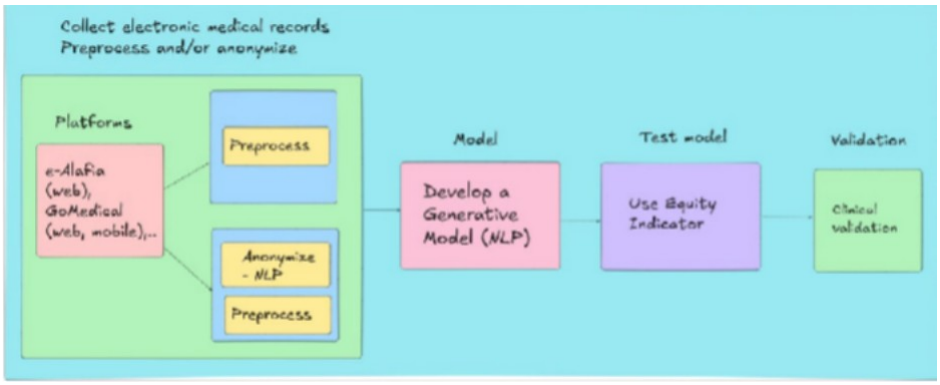

Figure 7: Fair NLP-based synthetic EMR generation pipeline in Benin

5.1. Description of the pipeline

Collection and Preprocessing of Electronic Medical Records (EMRs) The data are collected from platforms such as *e-Alafia* or *GoMedical* (web or mobile solutions used in Benin). Two key steps are involved:

- **Standard preprocessing:** data cleaning and structuring.
- **Anonymization using NLP techniques:** identifying and masking sensitive information such as names, addresses, and dates, without distorting the clinical content.

Development of the Generative Model (NLP) We use a medical text generation model (e.g., GPT, T5, MedPaLM) with the objective of generating realistic, non-identifiable synthetic EMRs. These synthetic records can then be used to train or test other clinical models without violating ethical or legal restrictions on patient data.

Equity-Based Evaluation A dedicated testing phase evaluates the *equity* of the generative model:

- Does it produce balanced records across gender, age, geographic region, or socioeconomic status?
- Are there signs of bias, omission, or overrepresentation?

We propose the use of **equity indicators** such as entity distribution, case diversity, and demographic balance to assess the fairness of the generated content.

Clinical Validation The final step involves human expert validation. The synthetic records are submitted to medical professionals for *clinical validation*, ensuring consistency, plausibility, and potential utility for training, testing, or research.

Original Contribution Unlike traditional approaches, this contribution structures the entire pipeline—from EMR collection to clinical validation—while placing **equity as a central evaluation criterion**. The proposed framework can be applied in contexts where real patient data are too sensitive to use directly, but where synthetic alternatives can support safe and effective model development.

5.2. Discussion and limitations of the performed bibliometric study

Analysis of the results of this study highlights the advances and challenges involved in using generative artificial intelligence models to synthesize Electronic Medical Records (EMRs). Recent publications show a marked growth in research on this subject, particularly between 2020 and 2024. Indeed, the approaches being explored are mainly based on Transformer- type architectures, deep learning, unsupervised learning, GPT and its variants. These models are distinguished by their ability to process large quantities of unstructured data, such as free text in medical records, and to generate coherent summaries tailored to clinical needs.

Key benefits include a significant reduction in the cognitive load on healthcare professionals, and faster decision-making thanks to clear, targeted summaries.

The use of generative AI models is transforming clinical practice by enabling rapid access to essential information, thereby reducing delays in decision-making. This improves the quality of care while minimizing human error. However, clinical adoption remains hampered by concerns about the transparency of models and their ability to adapt to specific cases. Models have yet to prove their reliability in complex and diverse environments, such as intensive care or rare diagnoses.

To improve the reliability, safety and acceptability of generative AI models, there are several avenues to explore:

Standardization of evaluations: clear evaluation frameworks need to be developed to compare the relevance and accuracy of models.

Bias reduction and model robustness: The integration of more diversified databases and the use of federated learning techniques can help limit bias.

Data confidentiality: Advanced encryption solutions and anonymization techniques must be implemented to protect sensitive medical data.

Clinical adoption: Collaboration between researchers, healthcare professionals and regulators is essential to develop user-centered tools that meet ethical and regulatory requirements.

These directions offer promising prospects for integrating generative AI models into healthcare systems in an efficient and ethical way.

Despite the contributions of this study, certain limitations must be acknowledged. Firstly, the literature search was limited to the Scopus database, which may restrict coverage of relevant literature available in other databases such as PubMed, IEEE Xplore or Web of Science. The absence of qualitative evaluation or clinical case studies also restricts appreciation of the real impact of generative AI models in practical medical contexts. Finally, publication bias and

methodological variations between the studies analyzed may influence the overall synthesis, justifying caution in generalizing the results obtained. This bibliometric analysis could be supplemented in future work by a systematic review focusing on the best- performing models or concrete clinical use cases.

Conclusion

The application of generative artificial intelligence models to the synthesis of electronic medical records (EMRs) represents a major advance in healthcare information management. These technologies offer unique opportunities to lighten the cognitive load on healthcare professionals, speed up clinical decision- making and improve the quality of care. However, significant challenges remain, notably related to data bias, textual hallucination errors, and issues of confidentiality and clinical acceptability. The results of this research highlight the progress made in integrating generative models, such as GPT and others, but also underline the need to standardize assessment methods and enhance data security. The clinical adoption of these technologies will depend on the ability to resolve these challenges, based on multidisciplinary collaborations between researchers, clinicians and regulators. Finally, future directions must focus on developing more robust and transparent models, improving ethical practices, and educating healthcare professionals in the use of these tools. By overcoming these obstacles, generative AI models could sustainably transform the digital health landscape and become indispensable allies in EMR management.

Declaration on Generative AI

In preparing this work, the authors used X-GPT-4 for grammar and spelling checking. After using these tools/services, the authors reviewed and corrected the content as needed and take full responsibility for the content of the publication.

References

[1] K. Goodman, P. Yi, D. Morgan, Ai-generated clinical summaries require more than accuracy, JAMA Network (2024). doi:10.1001/jama.2024.0555.

[2] A. Myers, al, Ai can outperform humans in writing medical summaries, Stanford HAI (2024).

[3] Shing, al, Ai can outperform humans in writing medical summaries, arXiv (2023).

[4] Tan, al, Transformer-based approaches for medical document summarization, J. Biomed. Informatics (2023).

[5] Vishal, al, Challenges in medical data for generative ai models, J. Med. Informatics (2023).

[6] Lee, al, Ontology-driven medical record summarization, Health Tech (2023).

[7] Nguyen, al, Biases in medical data and their impact on ai model generalization, Health Data Science (2023).

[8] Simmons, al, Reducing hallucinations in medical text summarization with validation filters (2023).

[9] H. Xie, A. Cebulla, P. Bastani, M. Balasubramanian, Trends and patterns in electronic health record research (1991-2022): A bibliometric analysis of australian literature, Int J Environ Res Public Health, vol. 21, no. 3 (2022). doi:10.3390/ijerph21030361.

[10] A. Haris, Q. Aini, Bibliometric analysis of electronic medical records (emr) acceptance and adoption: Trends, insights, and future directions, Eman Research (2024). doi:10.25163/angiotherapy.859700.

[11] J. J. Jr., A. S. Jose, G. G. Ettaniyil, J. S, J. Jose, Mapping the landscape of electronic health records and health information exchange through bibliometric analysis and visualization, Cureus, Apr. 2024 (2024). doi:10.7759/cureus.59128.

[12] Y. Xie, Y. Zhai, G. Lu, Evolution of artificial intelligence in healthcare: a 30-year bibliometric study, Front Med (Lausanne), vol. 11 (2024). doi:10.3389/fmed.2024.1505692.
[13] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu3, P. Biancone1, Artificial intelligence applications in healthcare: A bibliometric and topic model-based analysis (2024).
[14] F. Chen, L. Wang, J. Hong, J. Jiang, L. Zhou, Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record- based models, May 01, 2024, Oxford University Press (2024). doi:10.1093/jamia/ocae060.
[15] A. K. Shaikh, S. M. Alhashmi, N. Khalique, A. M. Khedr, K. Raahemifar, S. Bukhari4, Bibliometric analysis on the adoption of artificial intelligence applications in the e-health sector, Jan. 01, 2023, SAGE Publications Inc (2023). doi:10.1177/20552076221149296.
[16] E. Özmen, B. Emir, The role of machine learning algorithms in sepsis diagnosis: A retrospective overview using bibliometric analysis, OSMANGAZİ JOURNAL OF MEDICINE, vol. 46, no. 6, Sep. 2024 (2024). doi:10.20515/otd.1532158.
[17] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu3, P. Biancone1, The role of artificial intelli- gence in healthcare: a structured literature review, BMC Med Inform Decis Mak, vol. 21, no. 1, Dec. 2021 (2021). doi:10.1186/s12911-021-01488-9.

[18] E. Asgari, J. Kaur, G. Nuredini, J. Balloch, A. M. Taylor, N. Sebire, R. Robinson, C. Peters, S. Sridharan, D. Pimenta, Impact of electronic health record use on cognitive load and burnout among clinicians: Narrative review, 2024, JMIR Publications Inc (2024). doi:10.1186/s12911-021-01488-9.

[19] T. B. Brown, B. Mann, N. Ryder, Language models are few-shot learners (2020). doi:10.48550/arXiv.2005.14165.
[20] O. AI, J. Achiam, Gpt-4 technical report (2023). doi:10.48550/arXiv.2303.08774.
[21] T. H. Kung, M. Cheatham, A. Medenilla, Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models (2023). doi:10.1371/journal.pdig.0000198.
[22] C. Su, Z. Xu, J. Pathak, F. Wang, Deep learning in mental health outcome research: a scoping review, Springer Nature (2020). doi:10.1038/s41398-020-0780-3.
[23] F. Liu, P. Demosthenes, Real-world data: a brief review of the methods, applications, challenges and opportunities, BioMed Central Ltd. (2022). doi:10.1186/s12874-022- 01768-6.
[24] J. Luo, M. Ye, C. Xiao, F. Ma, Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery (2020) 647–656. doi:10.1145/3394486.3403107.
[25] I. Landi, B. S. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, J. T. Dudley, C. Furlanello, R. Miotto, Deep representation learning of electronic health records to unlock patient stratification at scale, NPJ Digit Med (2020). doi:10.1038/s41746-020-0301-z.
[26] F. Xie, B. Chakraborty, M. E. H. Ong, B. A. Goldstein, N. Liu, Autoscore: A machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records, JMIR Med Inform (2020). doi:10.2196/21798.
[27] Y. Huang, A. Talwar, S. Chatterjee, R. R. Aparasu, Application of machine learning in predicting hospital readmissions: a scoping review of the literature, BMC Med Res Methodol (2021). doi:10. 1186/s12874-021-01284-z.
[28] G. Dhiman, S. Juneja, W. Viriyasitavat, H. Mohafez, M. Hadizadeh, M. A. Islam, I. E. Bayoumy, K. Gulati, A novel machine-learning-based hybrid cnn model for tumor identification in medical image processing, Sustainability (Switzerland) (2022). doi:10.3390/su14031447.
[29] P. Wang, T. Shi, C. K. Reddy, Text-to-sql generation for question answering on electronic medical records, The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020, Association for Computing Machinery (2020) 350–361. doi:10.1145/3366423.3380120.

[30] S.-W. Cheng, C.-W. Chang, W.-J. Chang, H.-W. Wang, C.-S. Liang, T. Kishimoto, J. P.-C. Chang, J. S. Kuo, K.-P. Su, The now and future of chatgpt and gpt in psychiatry, John Wiley and Sons Inc (2023). doi:10.1111/pcn.13588.

[31] K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, K. P. Bennett, The problem of fairness in synthetic healthcare data, Entropy (2021). doi:10.3390/e23091165.